SAVITRIBAI PHULE PUNE UNIVERSITY

Data Science & Big Data Analytics

A PROJECT REPORT ON

on

"WEATHER FORECASTING USING LOGISTIC REGRESSION"

Submitted to the

Savitribai Phule Pune University

In partial fulfilment for the award of the

Degree of Bachelor of Engineering

(COMPUTER ENGINEERING)

Submitted By

Team Code:01

Team Members:

1. Bhushan Bhosale	Roll No:- 308
2. Sagar Gore	Roll No:- 328
3. Shivani Mangrule	Roll No:- 359
4. Akanksha Pawar	Roll No:- 371

Under the guidance of

Prof. Nimbalkar. S. N

DEPARTMENT OF COMPUTER ENGINEERING

SVPM's COLLEGE OF ENGINEERING

MALEGAON(Bk.), BARAMATI,

PUNE-413115.

2021 -2022

CERTIFICATE

This is to certify that

1. Bhushan Bhosale	Roll No:- 308
2. Sagar Gore	Roll No:- 328
3. Shivani Mangrule	Roll No:- 359
4. Akanksha Pawar	Roll No:- 371

Group No. **01** Class **Third Year** Branch **Computer Engineering** has successfully completed thework associated with **data Science & Big Data Analytics(310256)** titled as

"WEATHER FORECASTING USING LOGISTIC REGRESSION"

and has submitted the work book associated under my supervision, in the partial fulfilment of Second Year Bachelor of Engineering (Choice Based Credit System) (2019 course) of Savitribai Phule Pune University.

Date: **18/4/2022**

Place: **MALEGAON (BK)**

Guide	Head	Principal
(Prof. Mhaske V. D)	(Prof. Kumbhar H. R.)	Dr. S. M. Mukane
(Prof. Kate-Deshmukh P. N.)		

ACKNOWLEDGMENT

We have great pleasure in submitting the project report on the topic

"WEATHER FORECASTING USING LOGISTIC REGRESSION"

First and foremost, we would like to thank All, for giving us the strength to complete the research project we started.

I would like to take this opportunity to thank my guides **Prof. Nimbalkar.S.N** the instructor for coaching & sharing ideas & real-life examples throughout the course. Without their guidance and support, this report could not be completed

We are also grateful to **Prof. Dabhade.S.A.** Head of Computer Engineering Department, **SVPM's College of Engineering** for his indispensable support, suggestions.

We are grateful to Principal **Dr. Mukane.S.M for** his encouragement and for giving us guidance and valuable time for the discussion. We express our sincere thanks toall the staffs, parents and colleagues who have helped us directly in completing thisproject. We are grateful for the many useful comments and suggestions provided by reviewers, which have resulted significant improvements in project

The five members of our group were a source of encouragement throughout, and we pitched it with an appreciable effort Lastly, we wish to express our sincere gratitude.

1. Bhushan Bhosale Roll No:- 308

2. Sagar Gore Roll No:- 328

3. Shivani Mangrule Roll No:- 359

4. Akanksha Pawar Roll No:- 371

INDEX

INTRODUCTION

Weather forecasting is the application of current technology and science to predict the state of the atmosphere for a future time and a given location. Different algorithms like Neural Network, Sliding Window, Bayesian, Decision Tree, etc. have been used to develop such forecasting applications in today's era. Most of the existing applications use data from the satellites, sensors and weather stations made for capturing weather data. Then, they analyze and predict weather with the help of computer models.

But in this project looking at the most affordable means of data collection and prediction, past 23 years daily weather data have been collected from Department of Meteorology and Hydrology, Naxal. Linear and Logistic regression algorithms have been used to predict the weather on the basis of minimum and maximum temperatures, wind speed, humidity, precipitation and chances of rainfall.Weather forecasting is the application of current technology and science to predict the state of the atmosphere for a future time and a given location.

Weather forecasts are made by collecting as much data as possible about the current state of the atmosphere (particularly the temperature, humidity and wind) and using understanding of atmospheric processes (through meteorology) to determine how the atmosphere evolves in the future.

However, the chaotic nature of the atmosphere and incomplete understanding of the processes mean that forecasts become less accurate as the range of the forecast increases. Traditional observations made at the surface of atmospheric pressure, temperature, wind speed, wind direction, humidity, precipitation are collected routinely from trained observers, automatic weather stations or buoys.

During the data assimilation process, information gained from the observations is used in conjunction with a numerical model's most recent forecast for the time that observations were made to produce the meteorological analysis.

Numerical weather prediction models are computer simulations of the atmosphere. They take the analysis as the starting point and evolve the state of the atmosphere forward in time using understanding of physics and fluid dynamics.

PROBLEM STATEMENT

" It is important to exactly determine the rain in Australia dataset contains about 10 years daily weather observations from numerous Australian weather stations. Here's a small sample from the dataset".

Problems concern availability, timeliness, and quality of observational data; time constraints on forecast preparation; the nature and reliability of communication systems available for forecast dissemination; and the makeup and requirements of the user community.

MOTIVATION

- The mission of the NWS is to reduce the loss of life and the loss of property associated with weather related hazards, and to mitigate the economic impact of disruptive weather.
- Weather forecasters look at current state of the weather and forecast maps and add their personal experience to come up with a forecast and to issue warnings.

OBJECTIVES

There are several reasons why weather forecast are important.They would certainty be missed if they were not there.It is a product of science that impacts on lives of many people.The following is a list of various reason why weather forecast are important:

1.Helps people prepare for how to dress(i.e. warm weather ,cold weather, windy weather, rainy weather)
2.Helps businesses and people plan for power production and how much power touse (i.e. power companies, where to set thermostat)
3. Helps people prepare if they need to take extra gear to prepare for the weather(i.e. umbrella, rain coat, sun screen)
4. Helps people plan outdoor activities (i.e. to see if rain/storms/cold weather willimpact outdoor event)
5. Helps curious people to know what sort of weather can be expected (i.e. a snowon the way, severe storms)
6. Helps businesses plan for transportation hazards that can result from the weather(i.e. fog, snow, ice, storms, clouds as it relates to driving and flying for example)
7. Helps people with health related issues to plan the day (i.e. allergies, asthma,heat stress)
8. Helps businesses and people plan for severe weather and other weather hazards(lightning, hail, tornadoes, hurricanes, ice storms)
9. Helps farmers and gardeners plan for crop irrigation and protection (irrigationscheduling, freeze protection)1.5 Problem and Existing Technology

METHODOLOGY/ PROPOSED SYSTEM BLOCK DIAGRAM

User will enter current temperature; humidity and wind, System will take this parameters and will predict weather from previous data in database. The role of the admin is to add previous weather data in database, so that system will calculate weather based on these data. Weather forecasting system takes parameters such as temperature, humidity, and wind and will forecast weather based on previous record therefore this prediction will prove reliable.

As the project encompasses different technical fields and requires specific handling for each subsection, the project is divided into six different sections

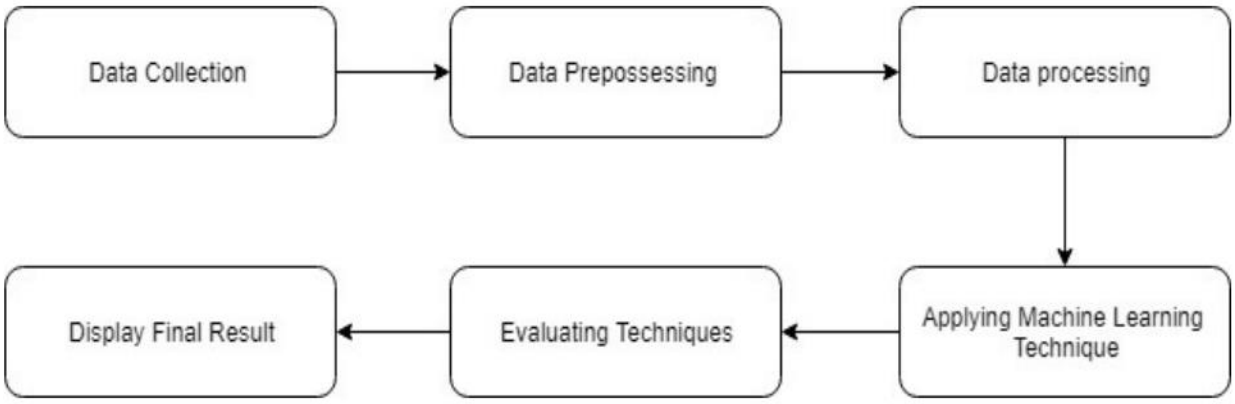


Figure 1: System Architecture

Features of Weather Forecasting

In future, we have plans to utilize low-cost Internet of Things (IoT) devices, such as temperature and humidity sensors, in collecting weather data from different parts of a city. The use of different sensors could increase the number of local features in the training dataset. This data, along with the weather station data, will further improve the performance of our prediction models. As machine learning advances, more models start integrating, and going with it accuracy becomes more better and forecasting will become increasingly accurate. More scope and potential of global nowcasting, which is relatively new addition to weather forecasting. As smart system penetration grows worldwide, more people will gain access to accurate, hyperlocal weather forecasting as well.

System Algorithm

Logistic regression

Logistic regression is used when the output are in categorical form. In this project, logistic regression has been used for forecasting the probability of rainfall which in turn decides whether it will rain or not.

Major objective of Logistic Function:

For this project, logistic regression is being for the categorizing the data and to predict the probability of rainfall. Major formulae used in the project:

The fundamental equation of generalized linear model is:

$g(E(y)) = \theta_0+x_1 0_1+x_2 0_2+x_3 0_3+x_4 0_4+x_5 0_5+\cdots(1)$

Where, g() is the link function, E(y) is the expectation of target variable and is the linear predictor i.e. minimum and maximum temperature, humidity, precipitation and wind speed. The role of link function is to link the expectation of y to linear predictor. The cost function is use to predict the optimum value of θ0,θ1,θ2,θ3,θ4,θ5. For those values cost function has minimum value and the predicted line is best fit.

Logistic regression is only concerned about the probability of outcome dependent variable (success or failure). As described above, g() is the link function.This function is established using two things:Probability of Success (p) and Probability of Failure(1− p). To make the probability less than 1, we must divide g(E(y)) by a number greater than g(E(y)) This can simply be done by:

$L(x) = \exp(g(E(y)))/1+\exp(g(E(y)))$ (2)

Using the above formula L(x) can be written as:

$L(x) = e^y /e^1+e^y$ (3)

where,

$g(E(y)) = \theta_0 +x_1 0_1 +x_2 0_2 +x_3 0_3 +x_4 0_4 +x_5 0_5$

L(x) is the probability of success of rainfall.

This (3) is the Logit Function

If L(x)is the probability of success, 1 − L(x) will be the probability of failure which can

be written as:

$q = 1-L(x) = 1- L(x)=1-e^{y} /e^{1}+e^{y}$ (4)

Where q is the probability of failure

Let L(x) = p

On dividing, Equation 3 by 4, we get:

$p/ 1- p = e^{y}$ (5)

After taking log on both side, we get

$Log(p/ 1- p)= y$ (6)

Log(p /1−p) is the link function. Logarithmic transformation on the outcome variable allows us to model a non-linear association in a linear way.

After substituting value of y, we'll get:

$Log(p /1- p) = 0_0+x_1 0_1+x_2 0_2+x_3 0_3+x_4 0_4+x_5 0_5$ (7)

where x1, x2, x3, x4, and x5 are different variables.

SOFTWARE AND HARDWARE REQUIRED

❖ Software:

1. Operating System: - Windows.
2. Python IDE:- Google Colab Notebook
3. Latest Version of Python
4. Python Libraries

❖ Hardware:

1. Processor: - Intel core i3.
2. Monitor: - Monochrome or color.
3. Memory: - 4 GB RAM.
4. Printer: - Any type of printer
5. Operating System:- Ubuntu or Microsoft Windows

Modules Description

In this project we have Two modules

1) Data gathering and pre - processing.
2) Applying Algorithm for prediction .

1) In this module we first gather the data(dataset) for our prediction model. Data comes in all forms, most of it being very messy and unstructured. They rarely come ready to use. Datasets, large and small, come with a variety of issues- invalid fields, missing and additional values, and values that are informs different from the one we require. In order to bring it to workable or structured form, we need to "clean" our data, and make it ready to use. Some common cleaning includes parsing, converting to one-hot, removing unnecessary data, etc. In our case, our data has some days where some factors weren't recorded. And the rainfall in cm was marked as T if there was trace precipitation. Our algorithm requires numbers, so we can't work with alphabets popping up in our data. so we need to clean the data before applying it on our model.

2) Logistic regression is used when the output are in categorical form. In this project, logistic regression has been used for forecasting the probability of rainfall which in turn decides whether it will rain or not.
Logistic regression is only concerned about the probability of outcome dependent variable (success or failure). As described above, g() is the link function. This function is established using two things: Probability of Success (p) and Probability of Failure

Module Outcomes:

1) By the end of the first module the fully cleaned and useful data is available for the apply the algorithm for the prediction

2) By the end of the second module the actual prediction will be happen the outcome is the amount of rainfall in inches based upon the users input

IMPLEMENTATION AND RESULT

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import io
```

```
import matplotlib
matplotlib.rcParams['font.size']=14
matplotlib.rcParams['figure.figsize']=(12,6)
matplotlib.rcParams['figure.facecolor']='#00000000'
sns.set_style('darkgrid')
```

```
from google.colab import files
```

```
uploaded=files.upload()
```

```
Choose Files  No file chosen        Upload widget is only available when the cell has been ex

Saving weatherAUS.csv.zip to weatherAUS.csv.zip
```

```
plt.figure(figsize=(12,10))
corr=data.corr()
sns.heatmap(corr,annot=True)
plt.show();
```

```
rows,columns=data.shape
print("The Number of rows in the dataset is:",rows)
print("The Number of columns in the dataset is:",columns)
```

Python

```
The Number of rows in the dataset is: 142193
The Number of columns in the dataset is: 24
```

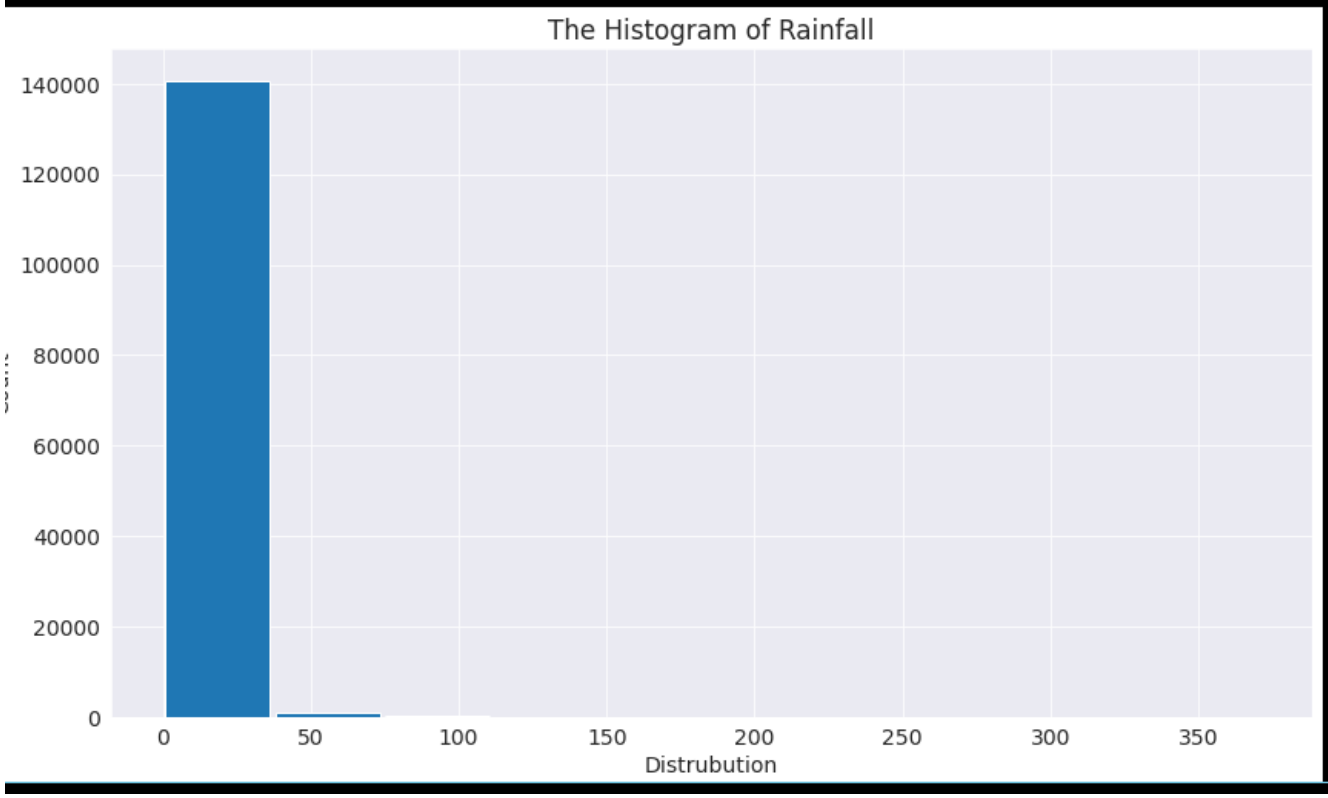
```
data.head()
```

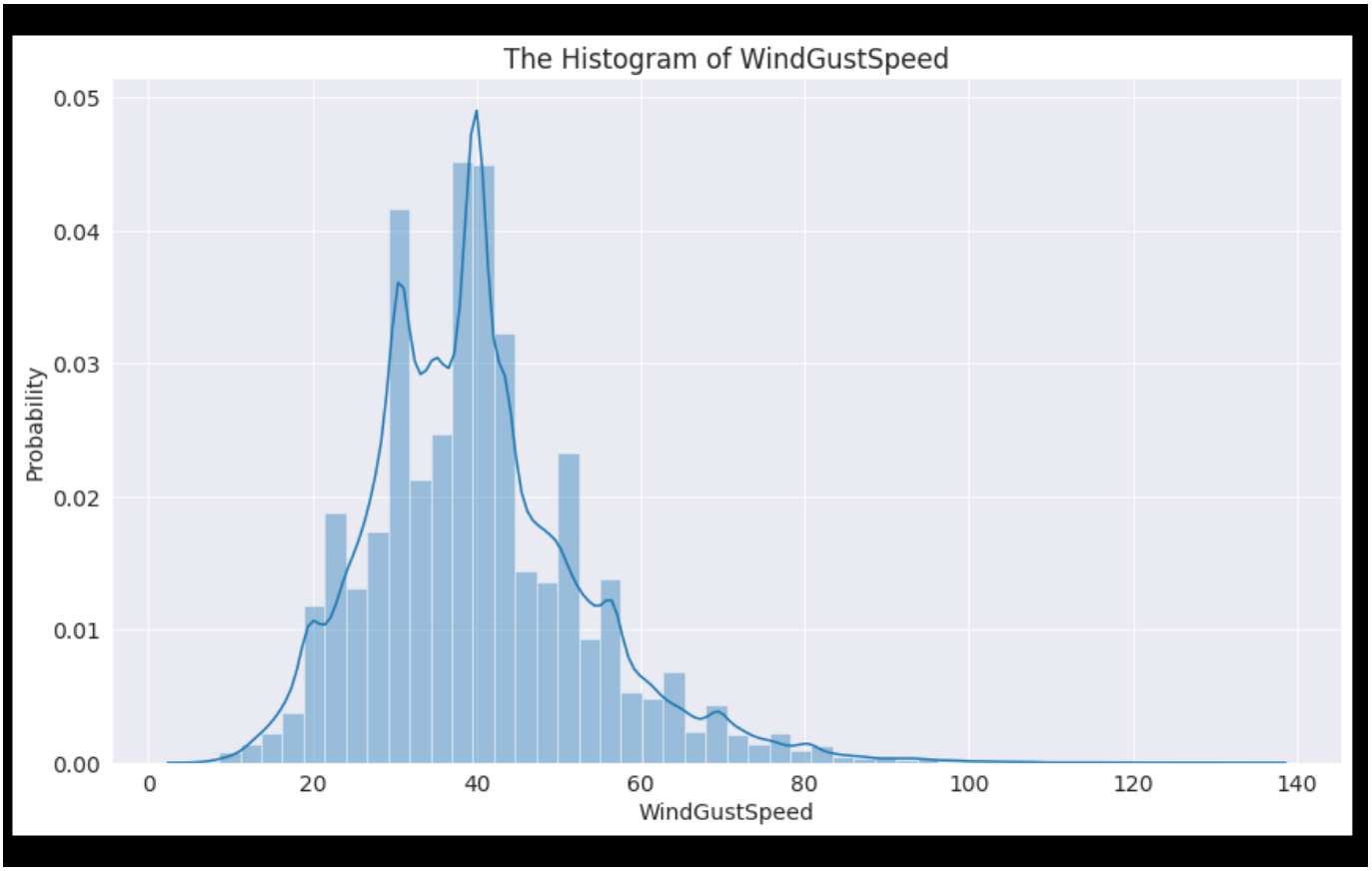
Python

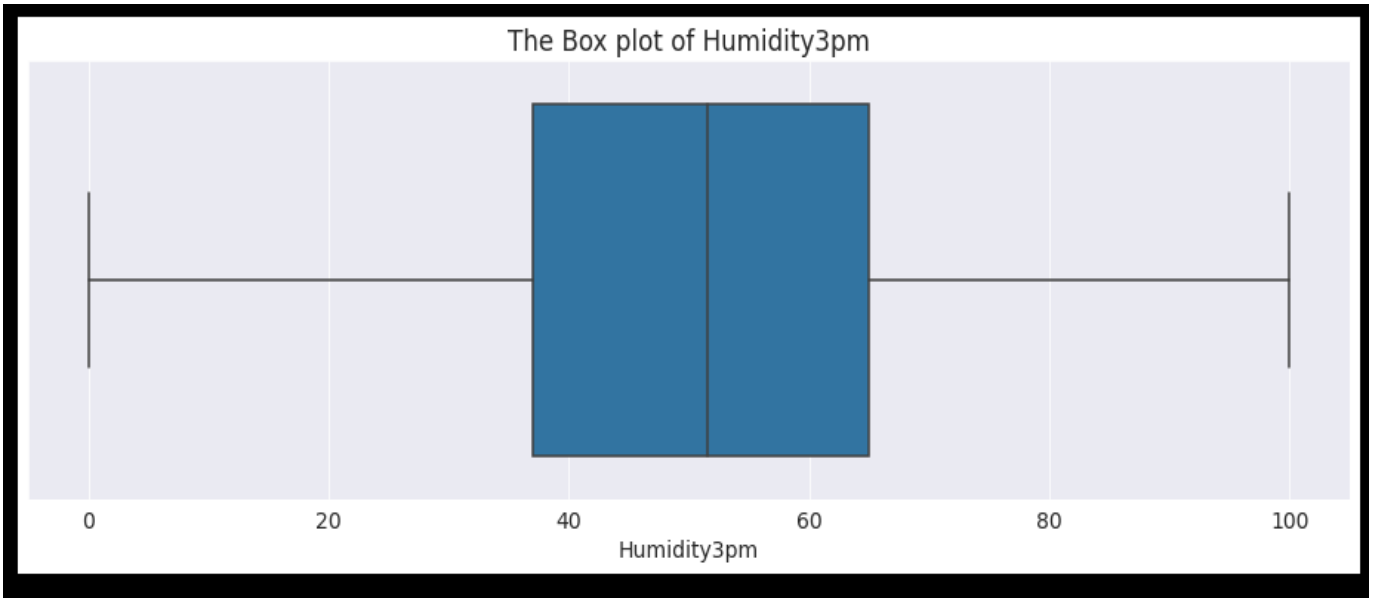
	Date	Location	MinTemp	MaxTemp	Rainfall	Evaporation	Sunshine	WindGustDir	WindGustSpeed	WindDir9am	...	Humidity3pm	Pressure9am	Pressure3pm	Cloud
0	2008-12-01	Albury	13.4	22.9	0.6	NaN	NaN	W	44.0	W	...	22.0	1007.7	1007.1	
1	2008-12-02	Albury	7.4	25.1	0.0	NaN	NaN	WNW	44.0	NNW	...	25.0	1010.6	1007.8	
2	2008-12-03	Albury	12.9	25.7	0.0	NaN	NaN	WSW	46.0	W	...	30.0	1007.6	1008.7	
3	2008-12-04	Albury	9.2	28.0	0.0	NaN	NaN	NE	24.0	SE	...	16.0	1017.6	1012.8	
4	2008-12-05	Albury	17.5	32.3	1.0	NaN	NaN	W	41.0	ENE	...	33.0	1010.8	1006.0	

5 rows × 24 columns

```
for column in Important_columns:
    Histogram(column)
```




The Histogram of WindGustSpeed
Probability
0.05
0.04
0.03
0.02
0.01
0.00
0
20
40
60
80
100
120
140
WindGustSpeed


The Box plot of Humidity3pm
0
20
40
60
80
100
Humidity3pm

Conclusion

We successfully predicted the rainfall using the linear regression but here this is not very accurate only some times any way it depends upon the climate changes to season to season. Here we are taking only summer season weather data set it only useful to predict rainfall in summer season

References

[1]Randy Olson. Accuracy of three major weather forecasting services. [Online]

http://www.randalolson.com/2014/06/21/accuracy-ofthree-major-weather-forecasting-services/, June 2014.

[2] Sam Savage. Firm expands ways to get weather. [Online] https://www.redorbit.com/news/technology/752863/firm expands ways to get weather/index.html/, December 2006. [Accessed: 03/06/2019].

[3] Jason Samenow. Accuweather extends its controversial,45-day weather forecasts to 90 days. [Online]

https://www.washingtonpost.com/news/capitalweather-gang/wp/2016/04/11/accuweather-extendsits-controversial-45-day-weather-forecasts-to-90-days/, April 2016. [Accessed: 03/06/2019].

[4] John Nese. Students put accuweather long-range forecasts to the test. [Online]

https://www.washingtonpost.com/news/capitalweather-gang/wp/2013/12/26/students-putaccuweather-long-range-forecasts-to-the-test/

[5] Jason Samenow. Accuweather: You cannot be serious(new 45-day forecasts). [Online]

https://www.washingtonpost.com/news/capitalweathergang/wp/2013/08/06/accuweather-youcannot-be-serious-new-45-day-forecasts/, August 2013. [Accessed03/06/2019].

[6] Nellie Andreeva and Mike Fleming Jr. Byronallen's entertainment studios acquires the

weather channel tv network for $300 million.

[Online] https://deadline.com/2018/03/byron-allenentertainment-studios-acquisition-the-weatherchannel-tv-network-1202351426/, March 2018.[Accessed: 03/06/201